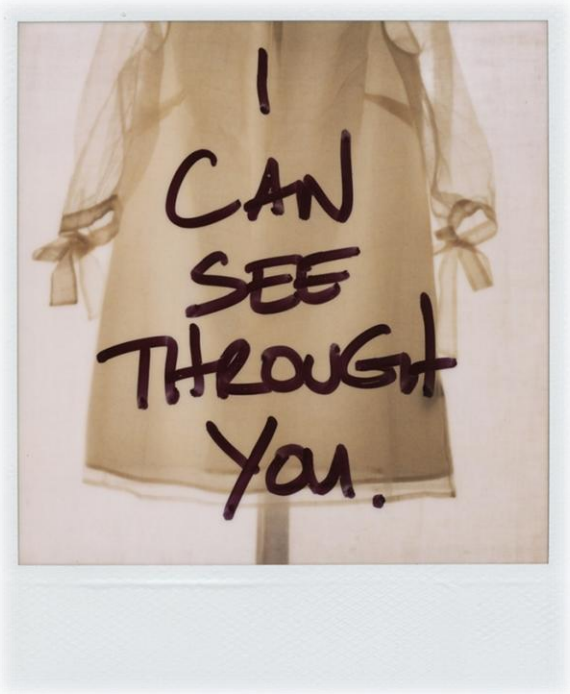
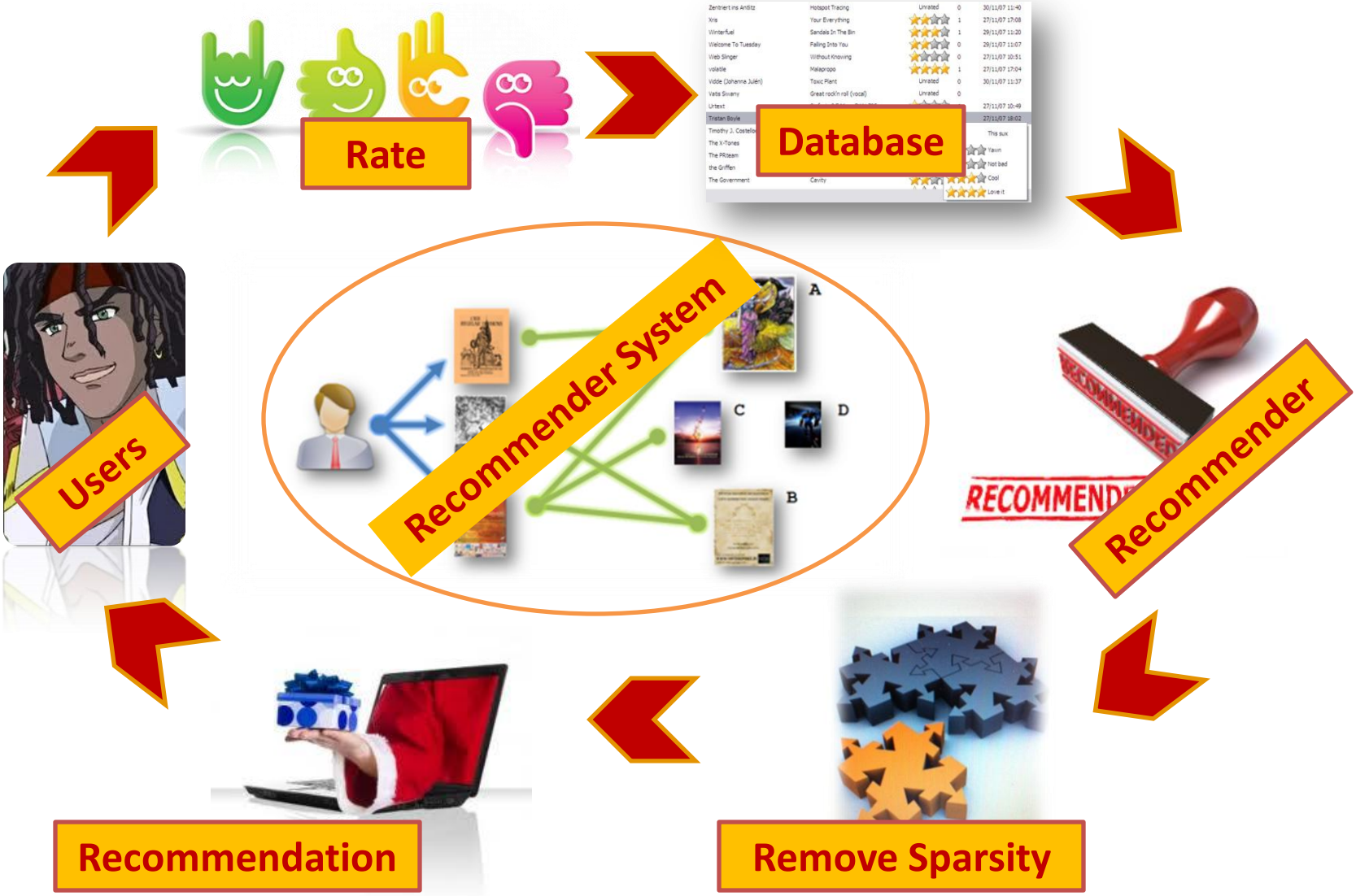
Attack Detection in Collaborative Filtering Recommender System



By:
Deepti Goel
Anurag Tripathi
Anvaya Rai

Architecture…



Four Pillars…

Recommender System…

"Automate the Circle of Advisors"

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	3	4	3	**?**	**?**
User 2	3	4	3	1	1
User 3	3	4	3	2	4
User 4	4	2	3	3	5
User 5	4	3	4	4	4
User 6	5	1	5	**?**	**?**

User 1 had already bought Item 1, 2 and 3.

Which Item should User 1 buy next, Item 4 or 5?

Average Rating of Item 4 is 10/4 = 2.5 and of Item 5 is 14/4 = 3.5

So, he should go for Item 5 !!!

Recommender System …

How correct was the previous recommendation ?

There could be many more parameters affecting the choice of the USER 1

Above recommendation would be same for USER 1 and USER 6

But there choices and needs may be different based on various other things like geographical location, culture, habits, liking etc

Taking these factors into account is the job of Collaborative Filtering during a recommendation!!!

Recommender System …

The basic idea…

- **Get the set of most similar users from the database**
- **Calculate the rating of the item by taking a weighted average , by giving more weights to the ratings of the similar users**
- **Make the recommendation**

Collaborative Filtering…

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?

Given User Parameters (Θ), Estimate Item Property Vectors (x) such that:

$$\min_{x^{(i)}} \frac{1}{2} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T(x^{(i)}) - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{k=1}^{n} (x_k^{(i)})^2$$

Where,

$r(i,j)$ = 1 if user j has rated movie I

$y^{(i,j)}$ = rating given by user j to movie i

The Threat… The Fake Profiles

True Database

Zentriert ins Antlitz	Hotspot Tracing	Unrated	0	30/11/07 11:40
Xris	Your Everything		1	27/11/07 17:08
Winterfuel	Sandals In The R		1	29/11/07 11:20
Welcome To Tuesday	Falling In			29/11/07 11:07
Web Slinger	W		0	27/11/07 10:51
volatile				27/11/07 17:04
Vidde (Johanna Julén)	nt	Unrated	0	30/11/07 11:37
tis Siwany	reat rock'n roll (vocal)	Unrated	0	
	Sinfonia 9 F Minor BWV 795		1	27/11/07 10:49
	After the Trip		1	27/11/07 18:02
Timothy J.	Ceilidh			
The X-Tones	FLIGHT of the FLY			
The PRteam	New Ideology			
the Griffen	Frail Signs			
The Government	Cavity			

This sux
Yawn
Not bad

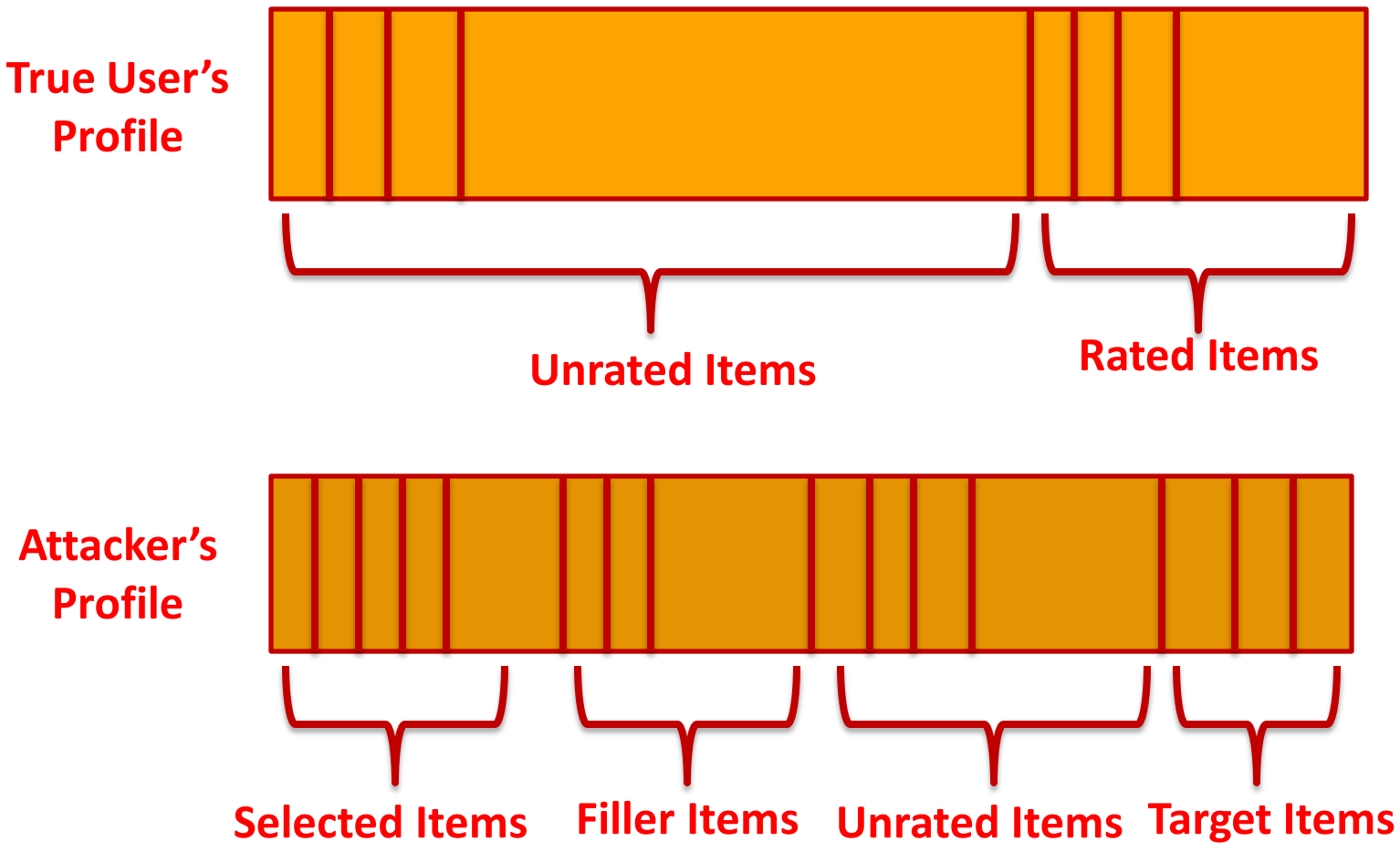
Fake Users

Distorted Database

Zentriert ins Antlitz	Hotspot Tracing	Unrated	0	30/11/07 11:40
Xris	Your Everything		1	27/11/07 17:08
Winterfuel	Sandals In The R		1	29/11/07 11:20
Welcome To Tuesday	Falling In		0	29/11/07 11:07
Web Slinger	W		0	27/11/07 10:51
volatile			1	27/11/07 17:04
dde (Johanna Julén)	lant	Unrated	0	30/11/07 11:37
Vatis Siwany	reat rock'n roll (vocal)	Unrated	0	
Urtext	Sinfonia 9 F Minor BWV 795		1	27/11/07 10:49
Tristan Boyle	After the Trip		1	27/11/07 18:02
Timoth	Ceilidh			
The X-Tone	FLIGHT of the FLY			
The PRteam	New Ideology			
the Griffen	Frail Signs			
The Government	Cavity			

This sux
Yawn
Not bad
Cool
Love it

True User and Attacker Profile…



The Attack Models…

Random Attack

Average Attack

Bandwagon Attack

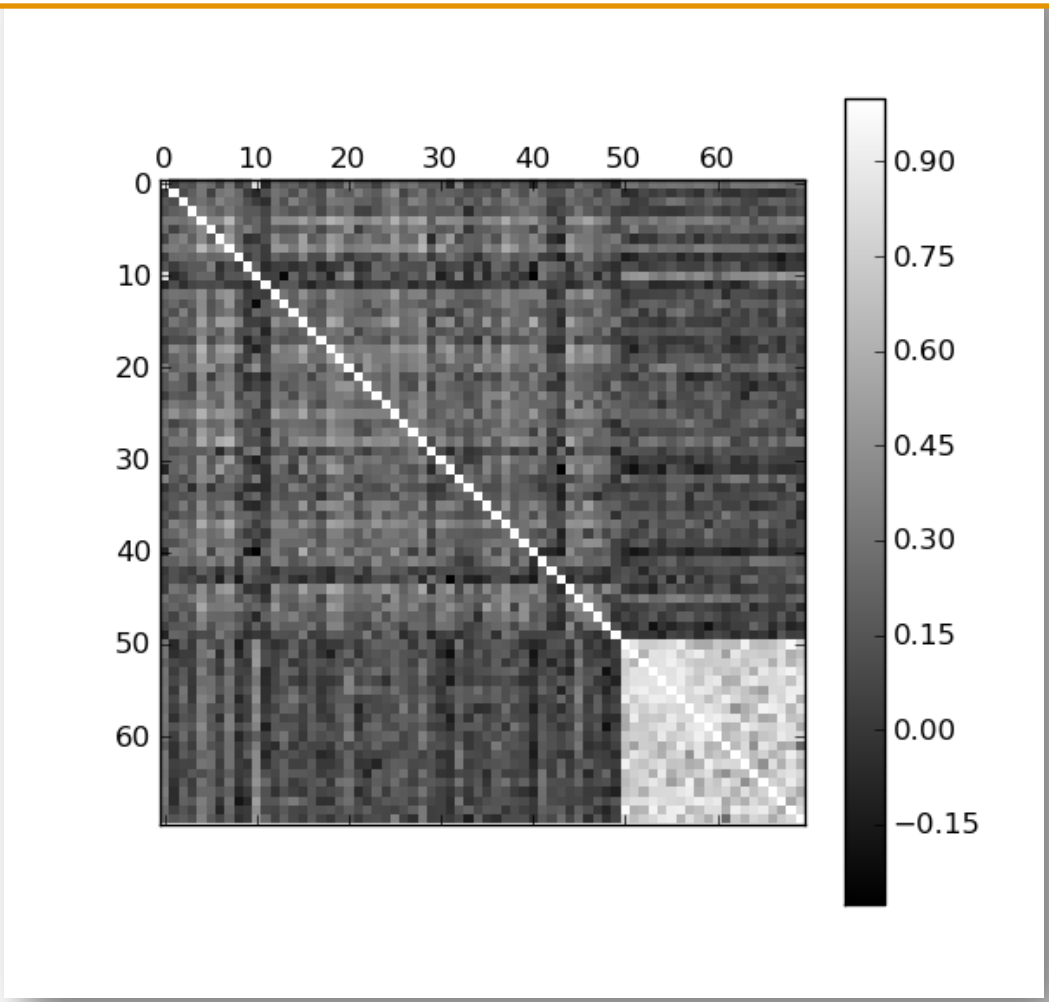
Segment Attack

Love-Hate Attack

Reverse Bandwagon Attack

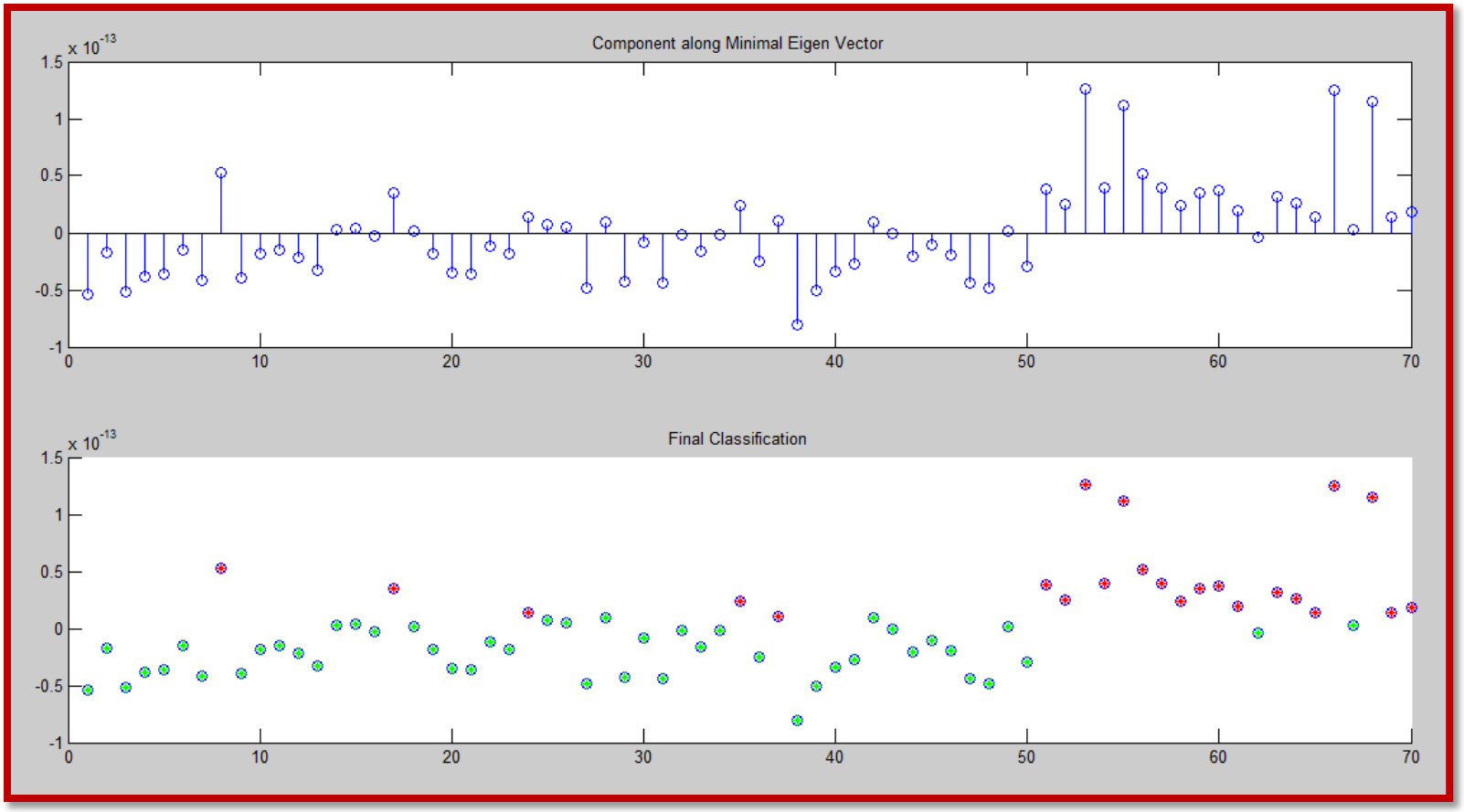
Attack Profiles created by these models effective in promoting an item, but they are highly correlated and hence can be detected by the Recommender System easily

The Correlation Matrix…



True Profiles have huge Variance but low Covariance and in case of Fake Profiles it is vice versa

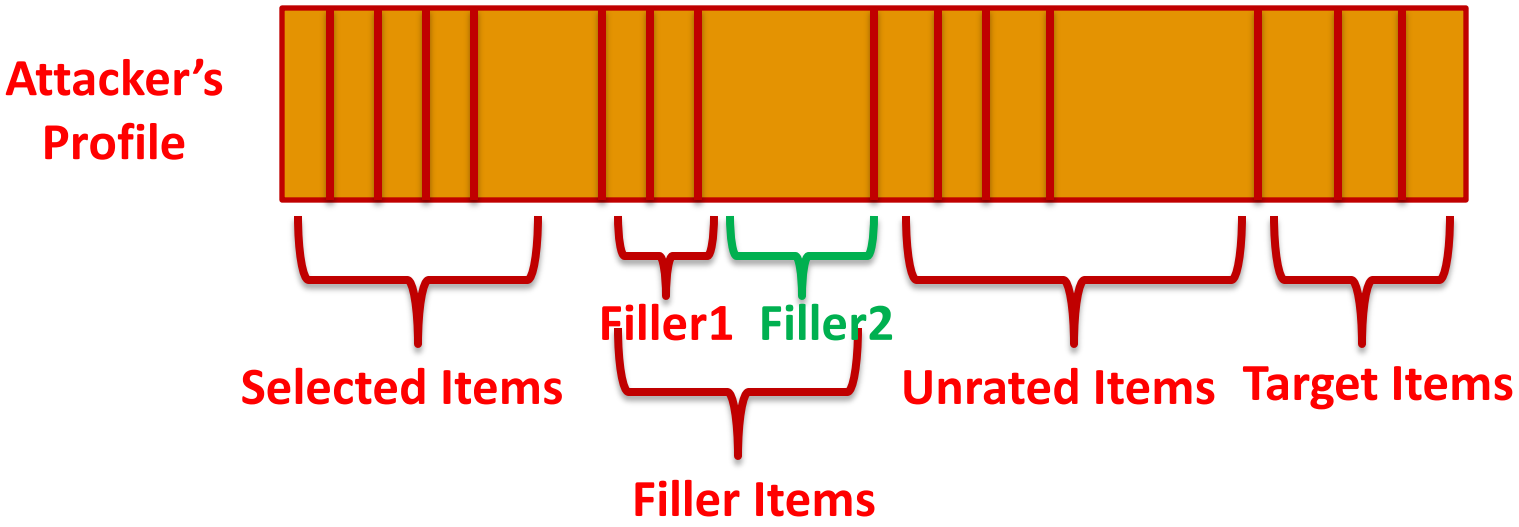
The Fake Profile Detection…



True Profiles in Green and Fake Profiles in Red

Detection done using PCA

The Hybrid Attack Model…

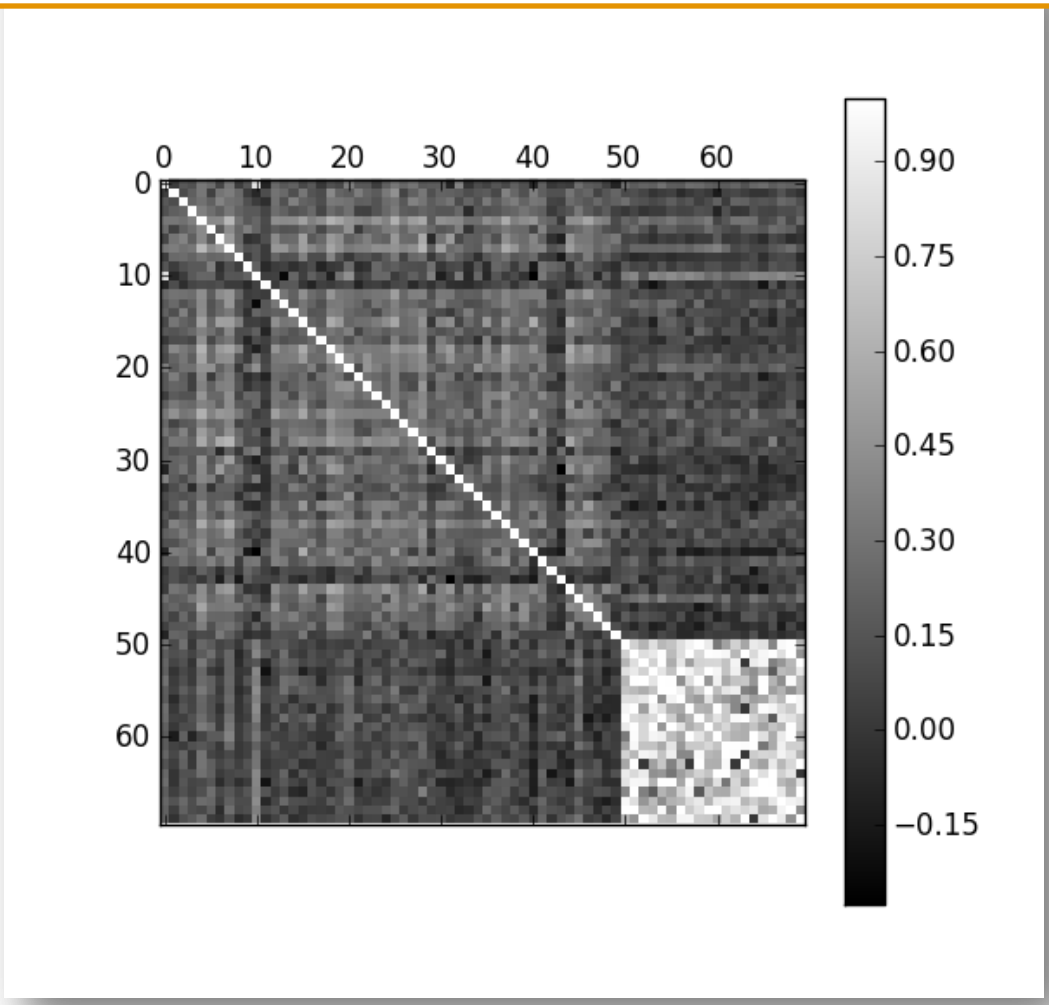


Filler Items split into two sets: Filler1 and Filler2

Filler1 filled using Gaussian with mean and variance of first 'F1' items in the list
Filler2 filled using Gaussian with mean and variance of first 'F1' items in the list

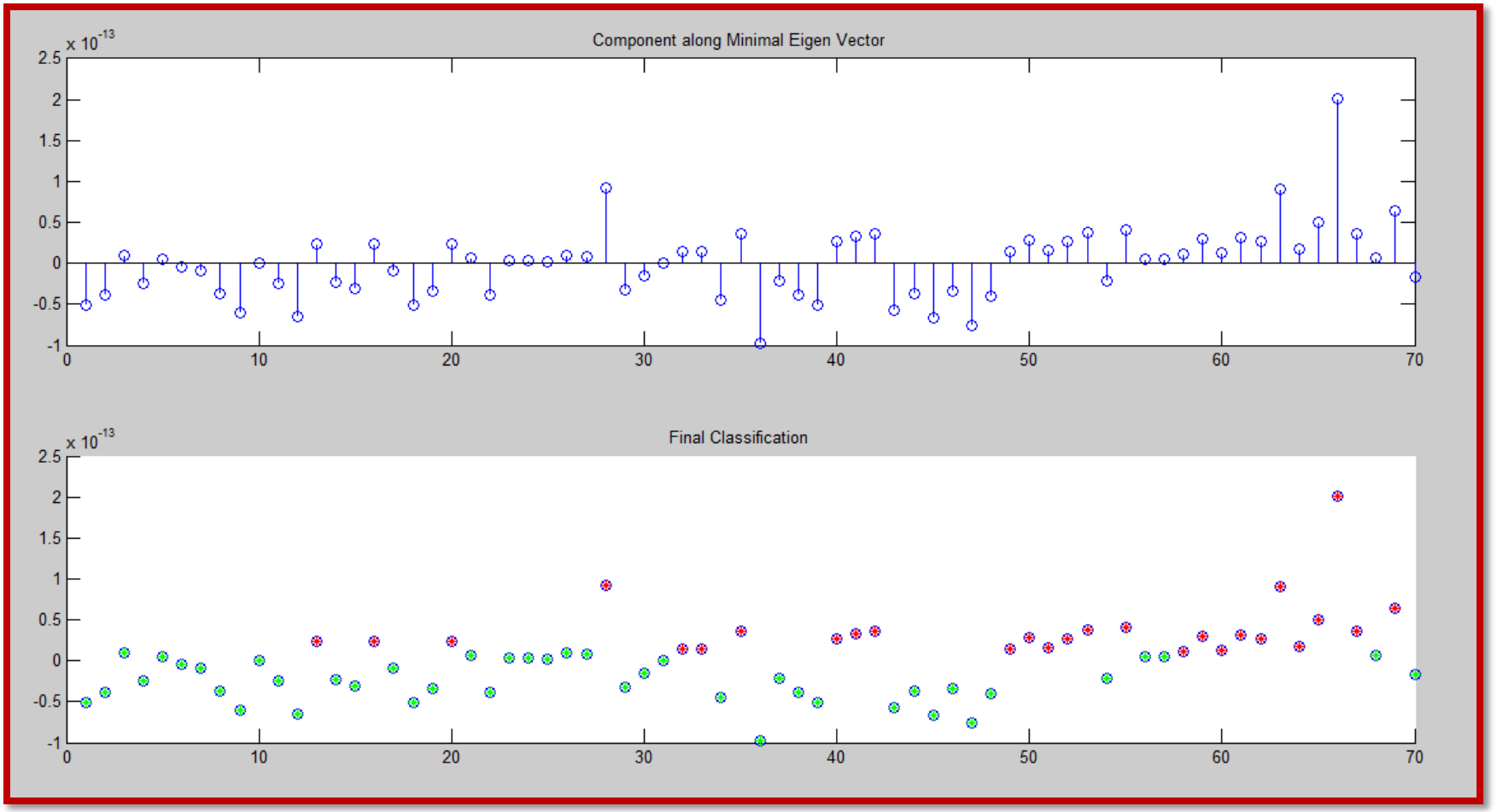
Items selected randomly and ratings assigned through Gaussian

The New Correlation Matrix…



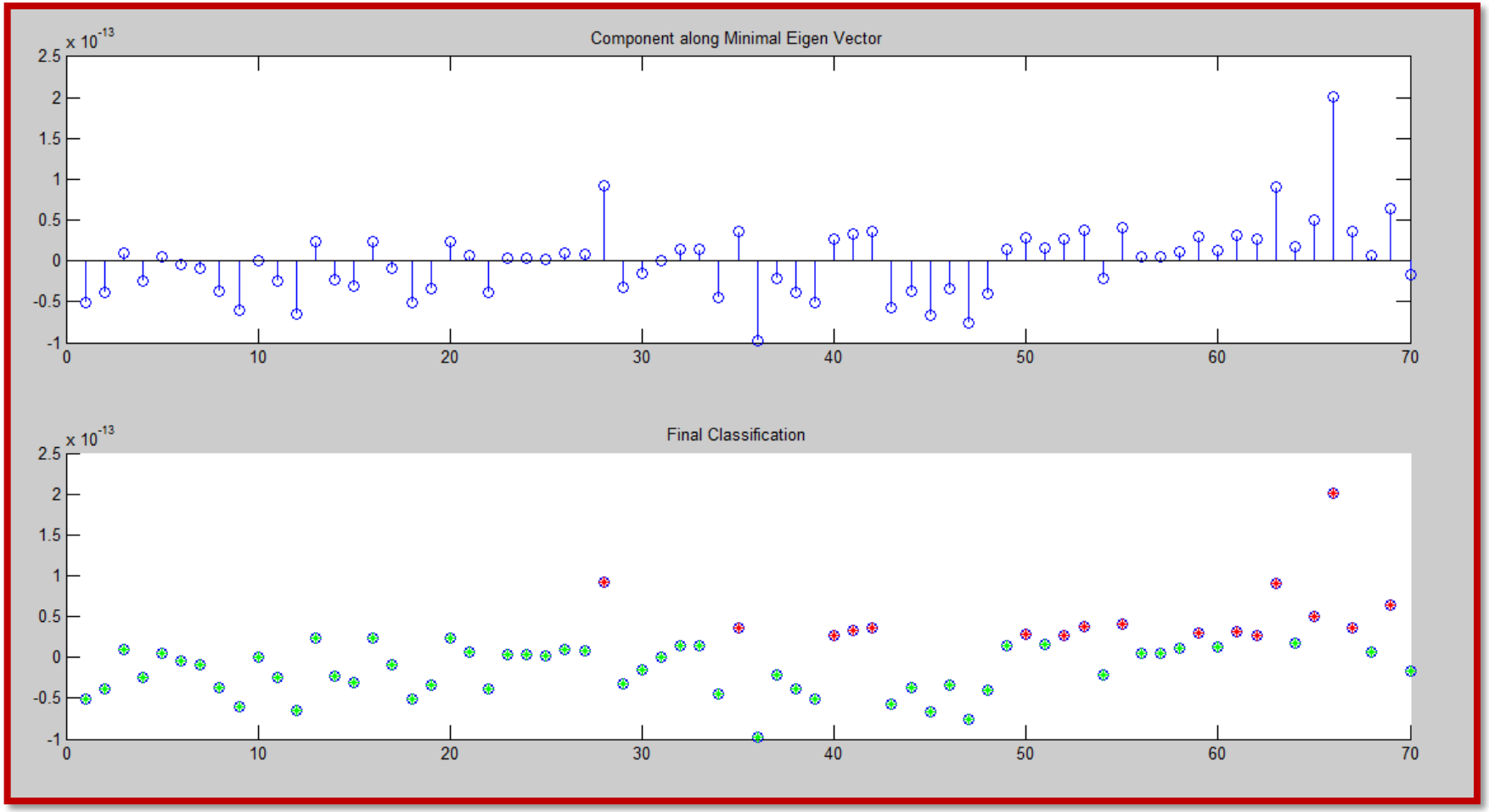
A reduced amount of correlation is visible among the Fake Profiles

The PCA Analysis…



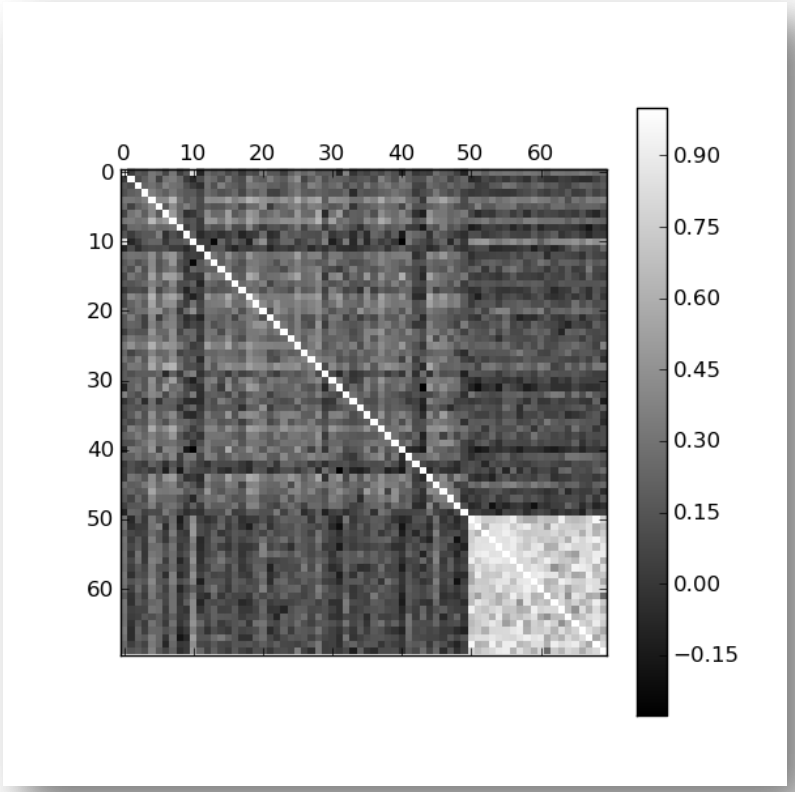
Poor classification with previous threshold

The PCA Analysis…

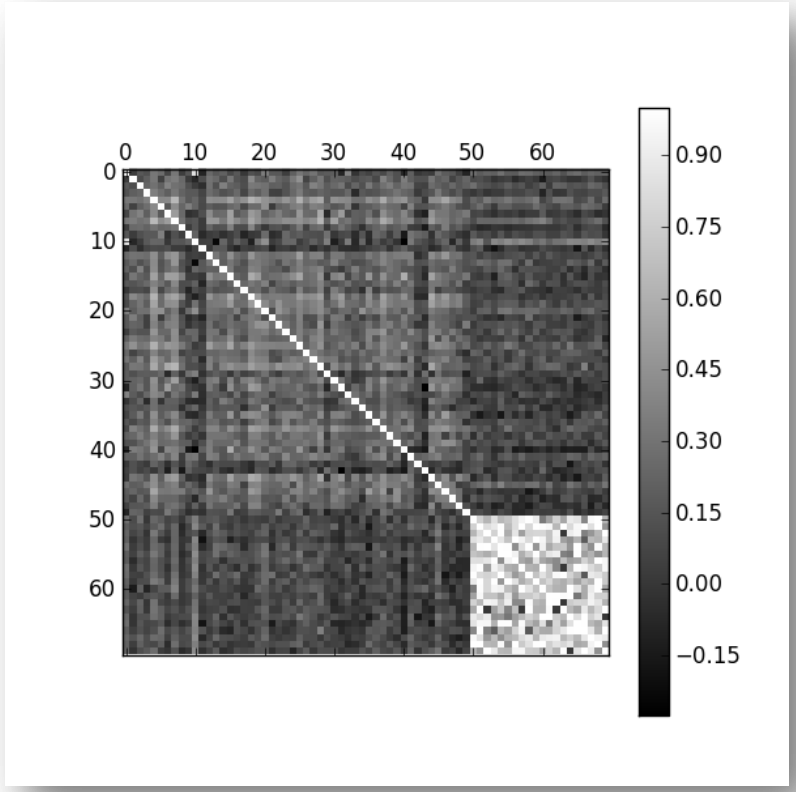


Poor classification with Updated threshold as well

The Results…



Bandwagon Attack Model



Hybrid Attack Model

Reduced correlation between the Fake Profiles

The Results…

Case	PCA		SVM	
	Bandwagon	Hybrid	Bandwagon	Hybrid
True Profiles classified as Fake Profiles	5	12	2	3
Fake Profiles classified as True Profiles	2	5	0	0
True Profiles classified as Fake Profiles (Threshold Updated)	NA	5	NA	NA
Fake Profiles classified as True Profiles(Threshold Updated)	NA	9	NA	NA

Increased Misclassification i.e. more True Profiles classified as Fake Profiles

The Resources…

Dataset:

MovieLens 100K data set (www.cs.umn.edu/research/GroupLens/data).

OS:

Windows and Linux

Platforms:

MATLAB and Python

Packages:

libSVM, scikit-learn and python-matplotlib

The References…

1.http://www.tud.ttu.ee/material/enn/Agent/attacking_recommender_systems.pdf

2. Recommender System Handbook, 2010

3. http://www.grouplens.org/

4. Unsupervised shilling detection for collaborative filtering http://dl.acm.org/citation.cfm?id=1619870

5. Attacks and Remedies in Collaborative Recommendation

Thank You!

Time for You to Attack…
The Questions?